Applied Mathematical Sciences, Vol. 7, 2013, no. 48, 2353 - 2356
HIKARI Ltd, www.m-hikari.com

The Shortest Common Superstring Problem

Anna Gorbenko

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
gorbenko.ann@gmail.com

Vladimir Popov

Department of Intelligent Systems and Robotics
Ural Federal University
620083 Ekaterinburg, Russia
Vladimir.Popov@usu.ru

Abstract

We consider the problem of the shortest common superstring. We describe an approach to solve the problem. This approach is based on an explicit reduction from the problem to the satisfiability problem.

Keywords: shortest common superstring, satisfiability, **NP**-complete

Investigation of different regularities can be used to retrieve various important knowledge (see e.g. [1] – [6]). In particular, different variants of the shortest common superstring problem play important roles in data compression and DNA sequencing. There are a number of variants of the shortest common superstring problem which are also of considerable interest for investigations.

Let U, V, X, and Y be some strings such that $U = XVY$. Then U is a superstring of V. The length of a string S is the number of letters in it and is denoted as $|S|$. Let $\Sigma = \{a_1, \ldots, a_m\}$ be a finite alphabet. Let $\mathcal{S} = \{S_1, \ldots, S_n\}$ be a collection of strings over Σ. A string S is a superstring

of $\mathcal{S}$ if S is a superstring of S_i, for any $1 \leq i \leq n$. The decision version of the shortest common superstring problem can be formulated as following.

The Shortest Common Superstring Problem (SCS):

Instance: *A collection $\mathcal{S}$ of strings over Σ and a positive integer k.*

Question: *Is there S such that $|S| \leq k$ and S is a superstring of $\mathcal{S}$?*

The problem SCS is **NP**-complete [7]. Encoding different hard problems as instances of different variants of the satisfiability problem and solving them with very efficient satisfiability algorithms has caused considerable interest (see e.g. [8] – [12]). We consider an explicit reduction from SCS to the satisfiability problem. Let

$$\begin{aligned}
\varphi[1] &= \wedge_{1\leq i\leq k} \vee_{1\leq j\leq m} x[i,j],\\
\varphi[2] &= \wedge_{1\leq i\leq k, 1\leq j[1]<j[2]\leq m}(\neg x[i,j[1]] \vee \neg x[i,j[2]]),\\
\varphi[3] &= \wedge_{1\leq i\leq n} \vee_{1\leq j\leq k-|S_i|+1} y[i,j],\\
\varphi[4] &= \wedge_{1\leq i\leq n, 1\leq j[1]<j[2]\leq k-|S_i|+1}(\neg y[i,j[1]] \vee \neg y[i,j[2]]),\\
\varphi[5] &= \wedge_{1\leq i\leq n, 1\leq j\leq k-|S_i|+1, j\leq s\leq j+|S_i|-1, 1\leq t\leq m, S_i[s-j+1]\neq a_t}(\neg y[i,j] \vee \neg x[s,t]).
\end{aligned}$$

Let $\xi = \wedge_{i=1}^{5}\varphi[i]$. It is easy to check that there is a string S such that $|S| \leq k$ and S is a superstring of $\mathcal{S}$ if and only if ξ is satisfiable. It is clear that ξ is a CNF. So, ξ gives us an explicit reduction from SCS to SAT. Now, using standard transformations (see e.g. [13]) we can obtain an explicit transformation ξ into ζ such that $\xi \Leftrightarrow \zeta$ and ζ is a 3-CNF. Clearly, ζ gives us an explicit reduction from SCS to 3SAT. We have designed generators of natural instances for SCS. We consider our genetic algorithms OA[1] (see [14]), OA[2] (see [15]), OA[3] (see [16]), and OA[4] (see [17]) for SAT. We used heterogeneous cluster. Each test was runned on a cluster of at least 100 nodes. Note that due to restrictions on computation time (20 hours) we used savepoints. Selected experimental results are given in Table 1.

time	OA[1]	OA[2]	OA[3]	OA[4]
average	3.44 h	3.78 h	3.63 h	2.42 h
max	21.16 h	27.12 h	26.57 h	18.04 h
best	16.02 min	17.07 min	16.59 min	8.41 min

Table 1: Experimental results for SCS.

ACKNOWLEDGEMENTS. The work was partially supported by Analytical Departmental Program "Developing the scientific potential of high school" 8.1616.2011.

References

[1] V. Yu. Popov, Computational complexity of problems related to DNA sequencing by hybridization, *Doklady Mathematics*, 72 (2005), 642-644.

[2] V. Popov, The approximate period problem for DNA alphabet, *Theoretical Computer Science*, 304 (2003), 443-447.

[3] V. Popov, The Approximate Period Problem, *IAENG International Journal of Computer Science*, 36 (2009), 268-274.

[4] V. Popov, Approximate Periods of Strings for Absolute Distances, *Applied Mathematical Sciences*, 6 (2012), 6713-6717.

[5] V. Popov, Multiple genome rearrangement by swaps and by element duplications, *Theoretical Computer Science*, 385 (2007), 115-126.

[6] V. Popov, Sorting by prefix reversals, *IAENG International Journal of Applied Mathematics*, 40 (2010), 247-250.

[7] J. Gallant, D. Maier, J.A. Storer, On Finding Minimal Length Superstrings, *Journal oj Computer and System Sciences*, 20 (1980), 50–58.

[8] A. Gorbenko and V. Popov, The Minimum Test Collection Problem, *Applied Mathematical Sciences*, 7 (2013), 1191-1193.

[9] A. Gorbenko and V. Popov, The Farthest Substring Problem, *Applied Mathematical Sciences*, 7 (2013), 1209-1212.

[10] A. Gorbenko and V. Popov, On Hamilton Paths in Grid Graphs, *Advanced Studies in Theoretical Physics*, 7 (2013), 127-130.

[11] A. Gorbenko and V. Popov, The Swap Common Superstring Problem, *Applied Mathematical Sciences*, 7 (2013), 609-614.

[12] A. Gorbenko and V. Popov, The String Barcoding Problem, *Applied Mathematical Sciences*, 7 (2013), 615-622.

[13] A. Gorbenko and V. Popov, The c-Fragment Longest Arc-Preserving Common Subsequence Problem, *IAENG International Journal of Computer Science*, 39 (2012), 231-238.

[14] A. Gorbenko and V. Popov, On the Problem of Placement of Visual Landmarks, *Applied Mathematical Sciences*, 6 (2012), 689-696.

[15] A. Gorbenko and V. Popov, Computational Experiments for the Problem of Selection of a Minimal Set of Visual Landmarks, *Applied Mathematical Sciences*, 6 (2012), 5775-5780.

[16] A. Gorbenko and V. Popov, Task-resource Scheduling Problem, *International Journal of Automation and Computing*, 9 (2012), 429-441.

[17] A. Gorbenko and V. Popov, SAT Solvers for the Problem of Sensor Placement, *Advanced Studies in Theoretical Physics*, 6 (2012), 1235-1238.

Received: February 12, 2013